

Variance-Based Sensitivity Analysis: An illustration  
on the Lorenz '63 Model

Caren Marzban<sup>1,2\*</sup>

<sup>1</sup> Applied Physics Laboratory, University of Washington, Seattle, WA 98195

<sup>2</sup> Department of Statistics, University of Washington, Seattle, WA 98195,

---

\*Corresponding Author: [marzban@stat.washington.edu](mailto:marzban@stat.washington.edu)

## Abstract

Sensitivity Analysis (SA) generally refers to an assessment of the sensitivity of the output(s) of some complex model with respect to changes in the input(s). Examples of inputs or outputs include initial state variables, parameters of a numerical model, or state variables at some future time. SA can be useful for data assimilation, model tuning, and dimensionality reduction. However, many commonly employed implementations of SA involve inadequate practices, such as simply varying the inputs and observing the effect on the outputs. This paper reviews some of the latest SA techniques, pointing out their pros and cons. Central to the discussion is the experimental design. Two methods are considered for sampling the input values: simple random, and space-filling designs, and it is shown that different measures of sensitivity have different properties under these sampling schemes. The methodology is illustrated on a few toy examples as well as on the Lorenz '63 model.

# 1 Introduction

In contemporary times it is commonplace to represent complex systems with numerical models. Examples include numerical weather prediction models (Richardson 2007), ocean circulation models (Miller 2007), and hydrology models (Rushton 2003). All of these models generally consist of a system of partial differential equations which are numerically integrated, subject to boundary and initial conditions, usually based on observations. Generally, such complex models can be viewed as a “black box” with some number of inputs and outputs. Although the choice of the inputs and outputs depends on the specific problem at hand, for the sake of concreteness in this paper they can be thought of as model parameters and forecasts parameters, respectively. The general situation is displayed in Figure 1.

Experiments involving numerical models generate data which have no experimental error. In other words, a unique set of values for the inputs will always produce the same output. Such data are called *computer data*, and experiments involving computer data are often called *in silico* - in contrast to *in vitro* or *in vivo* experiments performed, respectively, in a laboratory tube or in a living body. Again, the distinguishing characteristic of computer experiments is the absence of experimental error (Fang, Li, and Sudjianto 2006; Santner, Williams, and Notz 2003). Consequently, the analysis of computer data is somewhat different from that of “real” data. The framework for analyzing computer data is well-established (Santer et al. 2003). The introduction of the field in meteorological circles dates back to 1993 (Bowman, Sacks,

and Chang 1993). In spite of the nearly two decades since that work, most published works on analysis of computer data in meteorology appear to neglect that framework.<sup>1</sup>

Sensitivity Analysis (SA) is performed for a variety of reasons. At the most fundamental level, the identification of the inputs which affect the outputs may be for the sake of building a theory which relates the two sets of variables. By contrast, engineers are interested in inputs which do not affect the output, because then the system can be said to be robust with respect to those inputs. In other situations, the purpose of SA may be to rank the inputs in some order of importance, which itself may be used for the purpose of dimensionality reduction. Reducing the number of inputs may be a necessity for practical reasons (e.g., the cost of collecting data on inputs), or for statistical reasons (e.g., assuring that the number of cases available is considerably larger than the number of inputs). Another reason for performing SA is model tuning. Numerical models often have a large number of parameters whose values are not unambiguously known or even knowable. It is, therefore, useful to know which parameters have an affect on the outputs, and what is the nature of that effect. These different applications of SA are not necessarily mutually exclusive, but this paper deals mostly with the latter.

Methods for performing SA are widely varied (Bolado-Lavin and Badea 2008; Fang, Li, and Sudjianto 2006; Oakley and O'Hagan 2004; Saltelli et al. 2008, 2010;

---

<sup>1</sup>At the time of writing this article, according to the Web of Science citation index, Bowman et al. (2003) has been cited only 13 times. Of those, only 4 are from journals of the American Meteorological Society, namely Journal of Climate (3), and Monthly Weather Review (1). The rest are statistics and engineering journals.

Santner, Williams, and Notz 2003; Sobol' 1993). One of the simplest methods involves the systematic inclusion (or exclusion) of inputs while the performance of the model is monitored. If the inclusion of some input is found to not affect performance (e.g., mean squared error), then the outputs are considered to be insensitive to that input. As intuitive as this approach is, it has several defects, the most significant of which is that it does not allow for the identification of interactions between the inputs - the situation where the effect of the inclusion/exclusion of one input may depend on the inclusion/exclusion of some other input. Another issue that makes this approach impractical is that the number of possible ways in which one, or a set, of inputs can be included or excluded from the analysis is simply too large for this brute-force method to be feasible. Other problems are discussed in the next section.

Scatterplots of an output vs. inputs are also useful in SA. But, these, too, ignore interactions. It is possible to examine the interaction between two inputs, say  $x_1$  and  $x_2$ , by making a scatterplot of  $x_1$  vs.  $y_1$  and superimposing it on the scatterplot of  $x_2$  vs.  $y_1$ , where  $y_1$  is one of the outputs. If the two scatterplots display some unambiguous feature (e.g., a line), and if the two features are parallel, then one can conclude that there is no interaction between  $x_1$  and  $x_2$ . However, this method cannot be easily generalized to higher-order interactions (e.g., between 3 inputs). Also, each output must be considered one at a time.

It is worth mentioning that the question of how uncertainty in the inputs is propagated to the outputs is technically called Uncertainty Analysis (Cacuci 2003). Although there is some overlap between the techniques for performing SA and uncer-

tainty analysis, the questions addressed in the two methods are qualitatively different. Also different from SA is the question of the extent to which the introduction of a new observation affects the outputs. Note that “new observation” does not refer to a new input, rather to a new measurement made on an existing input. The current paper does not address either of these notions of SA. The focus is on the extent to which the outputs are affected by the various inputs across the full range of input values.

In statistics “The full range of input values” is called the experimental region, and the question of how to choose it is a topic of text books on experimental design (Douglas 2005). It is important in SA because the sensitivity of the outputs to the inputs can depend on how the experimental region is selected. The obvious choice of “every possible value” is usually unfeasible. Alternatively, one can take random samples from the experimental region. The notion of “random,” however, is ambiguous. Often, a sampling scheme is designed to optimize some quantity. For example, in estimating the mean of some quantity, one may aim to maximize the precision of the estimate. Different sampling schemes produce estimates with different precisions, and for some problems one can even derive exact/analytic results for identifying a unique sampling scheme which maximizes the precision of the mean. In this paper, after highlighting the role of precision in SA, two sampling schemes commonly employed in SA are reviewed. Although some exact/analytic results do exist (e.g., the theorem on the last page of McKay, Beckman, and Conover (1979)), the methods described here are illustrated only numerically.

In choosing a specific method for performing SA, it is important to realize that there is no single method that applies to all problems; all of the methods make some specific assumptions about the nature of the problem, and answer very specific questions. One class of SA methods with relatively broad applicability is closely related to the analysis of variance, arising in regression and experimental design problems. The specific question is How is output uncertainty (i.e., variance) apportioned among the inputs? This paper focuses on one of these so-called variance-based methods (also called *global* because they do not require a notion of a derivative).

SA involves several rather distinct ingredients. The first deals with the specific method for performing SA (e.g., derivative-based, variance-based). Then there is the issue of estimating certain mathematical quantities (e.g., mean, variance, conditional mean) from data; it is at this estimation stage where methods of experimental design enter the analysis. Some of the quantities which must be estimated are conditional expected values of an output, given some set of the inputs. The numerical estimation of these expected values is a large topic and constitutes another ingredient of SA. All of these are discussed in the following sections, followed by some examples. The main purpose of this article is to demonstrate one SA technique commonly employed in many fields, although not very commonly in meteorological circles<sup>2</sup>

---

<sup>2</sup>At the time of this article, a search for “sensitivity analysis” through the journals of the American Meteorological Society, returns 1331 articles. However, the inclusion of any of the following terms - representative of the approach discussed here - returns no relevant hits: global, variance-based, Saltelli, Santner, Sobol, Tarantola. The latter four are some of the peoples’ names generally associated with SA.

## 2 Methods for SA

Given the long history and intuitive nature of SA, there exists a wide range of methods for implementing it. As mentioned in the previous section, one of the more brute-force approaches involves monitoring the performance of the “black box” model while some set of the inputs is either included or excluded. In addition to the shortcomings mentioned there, another problem with such approaches is that the “forward” and “backward” approaches often yield different importance levels for a given input (Draper and Smith 1998). Also, the choice of which inputs to include or exclude, and in what order, is not unique. As such, the results of such SA approaches can be ambiguous.

In such a stepwise approach, either a single input or a set of inputs may be included or excluded. The former falls in a class of methods often referred to as “One-at-A-Time” (OAT) methods (Saltelli et al. 2008, 2010). Generally, in an OAT analysis the inputs are varied one at a time, while all other inputs are held fixed at some value (e.g., at their respective mean), and the change in the output is monitored. However, if the number of inputs is large, the basic OAT approach samples only a small portion of the experimental region. This can be seen as follows: Consider three inputs whose values vary along the  $x, y, z$  Cartesian coordinates. Varying them one-at-a-time, will sample the points along the axes, but not at the corners of a cube centered at the origin. In 3-dimensions this is not a serious concern because one often assumes that the “black box” model is some relatively smooth surface, so that knowledge of

its output values along the three axes is sufficient to define it uniquely. However, it is a geometric fact that a high-dimensional space consists mostly of corners (Jimenez and Landgrebe 1998; Scott 1992), and so the basic OAT severely under-samples the experimental region. There exists a variation on the basic OAT, proposed by Morris (1991), which avoids both of these problems, but it will not be discussed here.

The taxonomy of SA methods is not simple (Bolado-Lavin Badea 2008), but one can divide them into two broad categories: local and global. Local methods yield sensitivity results which are true only in a small region of the experimental region. The basic OAT approach is a local method because its sensitivity results are reliable only in the vicinity of the fixed values assigned to the inputs. Another characteristic of local methods is their dependence on the notion of a derivative (of the output with respect to an input). For example, in the basic derivative-based approach, one generally examines the rate of change in the output with respect to the specific input being changed, in the neighborhood of some specific point in the experimental region. Consequently, local approaches are generally incapable of addressing large changes or interactions between the inputs. By contrast, global methods allow not only an assessment of the importance of a given input across a large range, but also how each input interacts with other inputs, across the full experimental region.

Although global methods themselves can be subdivided into finer categories, they are generally based on some decomposition of the variance of the output(s). The decomposition is done in a way so as to allow identification of the various terms in the decomposition with the different inputs and their interactions. In this way, one

can assess the manner in which the uncertainty in an output is apportioned across the inputs, and across interactions between them. The variance-based methods are global in the sense that the sensitivity results do not pertain to any specific value of the inputs, while at the same time are capable of assessing interactions. This generality of variance-based SA methods is the reason why it is the main focus of this work.

## 2.1 Variance-Based SA

The specific formulation of variance-based SA presented here follows that of Oakley and O’Hagan (2003). The foundations of the variance-based approach are based on two theorems. The first is known by the name variance-decomposition formula (Weiss 2005, p. 385):

$$V[y] = V[E[y|x]] + E[V[y|x]] \quad , \quad (1)$$

where  $E[\cdot]$  and  $V[\cdot]$  denote expected value and variance, respectively.  $E[y|x]$  and  $V[y|x]$  denote the conditional expected value and conditional variance, respectively, of the output, given an input  $x$ . Intuitively, this decomposition states that the total variance in  $y$ ,  $V[y]$ , can be written as the sum of two terms, one measuring the variance “between” the conditional means, and the other measuring the mean of the conditional (“within”) variances. The two terms are also known as the “explained” and the “unexplained” variance.

The second theorem is often known as the High-Dimensional Model Representa-

tion (Sobol' 1993); it states that any function of the type,  $y = \eta(x_1, x_2, \dots, x_n)$ , can be decomposed as follows:

$$y = \eta(x_1, x_2, \dots, x_n) = E[y] + \sum_i^n z_i(x_i) + \sum_{i < j} z_{ij}(x_i, x_j) + \dots , \quad (2)$$

where

$$z_i(x_i) = E[y|x_i] - E[y] , \quad (3)$$

$$z_{ij}(x_i, x_j) = E[y|x_i, x_j] - E[y|x_i] - E[y|x_j] + E[y] . \quad (4)$$

The  $z_i(x_i)$  are referred to as main effects, and the  $z_{ij}(x_i, x_j)$  are called the first-order interactions. The ... indicates that there exist higher-order interaction terms in the expansion, but here they are assumed to be relatively small. It is important to point out that the computation of the expected values and variances requires the **joint probability distribution** of all the inputs. As a result, even a main effect computation for a given input generally involves the other inputs.

## 2.2 Variance-based Measures of Sensitivity

The measure of importance for an input is a user-dependent quantity, but a few common measures natural to the formulation of the variance-based methods are as follows. One natural measure gauges the expected reduction in the variance of the output, given an input:  $E[V[y] - V[y|x_i]]$ . By equation (1), this measure, denoted  $V_i$ , can be written as

$$V_i = V[E[y|x_i]] . \quad (5)$$

As mentioned previously, it is desirable to have some measure of the sensitivity of the output on some *set* of inputs. For example, one can measure the expected reduction in the variance of the output, given  $x_i$  and  $x_j$ . By a similar argument, that measure can be written as

$$V_{ij} = V[E[y|x_i, x_j]] \ . \quad (6)$$

Another useful quantity measures the uncertainty remaining in the output, given all inputs, except  $x_i$ . For example,

$$V_{T1} = V[y] - V[E[y|x_2, x_3, \dots]] \ , \quad (7)$$

measures the remaining/unexplained variance in  $y$  after all inputs have been fixed, except  $x_1$ .

Traditionally, the  $V_i$  and  $V_{T_i}$  measures are converted to proportions, as follows:

$$S_i = V_i/V[y] \quad \text{and} \quad S_{T_i} = V_{T_i}/V[y] \ , \quad (8)$$

where  $S_i$  and  $S_{T_i}$  are called the main effect index, and the total effect index, respectively. Although they measure different facets of the importance of the  $i^{th}$  input, they do not capture interactions between the inputs. Therefore, it is important to supplement  $S_i$  and  $S_{T_i}$ , with  $V_{ij}$  for a more complete assessment of sensitivity. Table 1 provides a summary of these measures and their meaning.

## 2.3 Two Examples

In order to become familiar with the variance-based SA approach, two examples are considered in which the various sensitivity measures can be computed analytically. A simple example (from Oakely and O’Hagan 2003) of the “black box” model is

$$y = \eta(x_1, x_2) = x_1 \quad . \quad (9)$$

As mentioned previously, the computation of the various sensitivity measures requires a specification of the probability distribution of the  $x_i$ . Without specifying that distribution, the results are shown in the middle column of Table 1. Some of the measures are sufficiently simple to be meaningful. For example,  $V_{T_2}$  and  $S_{T_2}$  are both zero, reflecting the absence of an  $x_2$  term in the function. For the same reason,  $S_1$  is 1. Many of the measures are not zero, when one might expect a zero. For instance, the absence of an  $x_2$  term in the function may suggest that  $V_2$  should be zero, but it is not. The reason for that is the existence of an underlying distribution over  $x_1, x_2$ . Because of that distribution, the two inputs are generally not independent, and therefore,  $E[x_1|x_2]$  is nonzero. In short, the structure of the function  $\eta(\cdot)$  is not the only determinant of the sensitivity measures; the probability distribution of  $x_1, x_2$  is also a contributing factor.

If  $x_1, x_2$  are independent, then their probability distribution can be written as a product of the marginal distributions for each. Then,  $E[x_1|x_2] = 0$ , and the sensitivity measures simplify to quantities which do reflect the structure of the function itself, as expected. These are shown in the right column of Table 1, where the only unspecified

quantity is the variance of the single input.

The second example is useful because it takes the form of a regression model with an interaction term. In particular, this example illustrates how the sensitivity measures are related to regression coefficients. The  $\eta(\cdot)$  model is

$$y = \eta(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 . \quad (10)$$

In the general case where the probability distribution for  $x_1, x_2$  is not specified, the results are unwieldy (not shown). Even for independent  $x_1, x_2$ , the results are complex (middle column of Table 2). Indeed, they are shown here only to highlight their complexity, and to point out that regression coefficients do not generally measure sensitivity. Note that the effect of the interaction term in the model can be traced through all of the measures by identifying  $\beta_{12}$ . In particular, it can be seen that an interaction term in the model effects all of the measures in Table 2. (Although  $V[y]$  is not explicitly expanded in terms of the  $\beta$ -coefficients, it too depends on  $\beta_{12}$ .)

Further simplification occurs if one assumes that the inputs have been centered (i.e.,  $E[x_i] = 0$ ). The results (right column in Table 2) are more revealing. For example,  $z_1, z_2$ , and  $z_{12}$  are simply the three terms in the model itself (Eq. 10).  $S_1$  and  $S_2$  can be identified as the standardized regression coefficients (Draper and Smith 1998), or Pearson's correlation coefficients (squared) between  $y$  and  $x_1$ , and between  $y$  and  $x_2$ , respectively. In other words, if the SA analysis is performed on inputs which are independent and centered, then  $S_1$  and  $S_2$  can be readily computed by performing multiple linear regression (without interaction) on  $x_1, x_2$  and  $y$ , with the

latter as the response variable. Interestingly, this conclusion is true even if the “black box” model contains interactions between the inputs. The reason for this is evident in the manner in which  $\beta_{12}$  is always multiplied by an  $E[x_i]$ ; see Table 2.

### 3 Estimation of $E[\text{output}|\text{input}]$

Note that all of the measures in Table 1 can be computed from the  $z_i$  and  $z_{ij}$  defined in equations (3) and (4), which are written in terms of conditional expected values. A great deal of the work in variance-based SA methods is focused on efficient and accurate ways for estimating these conditional expectations from data. The data themselves are generated by evaluating the function  $\eta(\cdot)$  for some set of  $x_i$  values. The choice of the  $x_i$  values is a complex issue belonging to the realm of experimental design, described in the next section.

The methods for estimating the conditional expectations are varied, but they can be broadly divided into Monte Carlo methods (Sobol’ 1993; Cukier et al. 1973; Saltelli et al. 2008, 2010), and methods based on emulation, or meta models (Dusby 2008, Oakley and O’Hagan 2004). In the former, the conditional expected values are expressed as their defining integral form, which are then evaluated using Monte Carlo techniques. The latter methods aim to develop a statistical model that approximates  $y = \eta(\cdot)$ . The resulting statistical model is said to emulate the “black box” model  $\eta(\cdot)$ . The emulator is then employed to estimate the conditional expectations. One advantage of the emulator approach (over direct Monte carlo) is that it injects

smoothing constraints into the estimation, leading to more robust results.

The main approach adopted in this paper is a simple emulation approach. Recall that an ordinary least-squares linear fit to data on a pair of variables  $(u, v)$  is designed so that a prediction made by the line estimates the conditional expected value of the response  $(v)$ , given the predictor  $(u)$ . In other words,  $E[v|u]$  can be estimated by simply using the equation of the least-square fit to data on  $v$  vs.  $u$ . Indeed, the linearity assumption is unnecessary;  $E[v|u]$  can be parametrized as any function of  $u$ . As long as there exists sufficient data to properly estimate the parameters of the statistical model (e.g., regression coefficients), the least squares fit will estimate the conditional expectation (Bishop 1996; pages 201-202).

The development of an emulator is a complex and sophisticated procedure. Oakley and O’Hagan (2003) develop a Gaussian Process emulator, whose mathematics is similar to Kriging (Chaloner and Verdinelli 1995; Sacks, Schiller, and Welch 1989; Sacks, et al. 1989; Welch et al. 1992). These methods are designed to produce accurate and precise approximations of the “black box” model but with the least amount of data possible. In spite of the straightforwardness of the emulation approach, the development of a realistic emulator is still a complex process, calling for mature statistical modeling. The focus of the current paper is sensitivity analysis, and not rigorous emulation. So, to avoid the potential difficulties associated with complex emulators, a simple polynomial regression model of order 2 is used to estimate the conditional expectations. Specifically,  $E[y|x_1]$  is estimated by fitting a curve of the form  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$  through data on  $x_1$  and  $y$ . Similarly for

$E[y|x_2]$ , etc.. The quantity  $E[y|x_1, x_2]$  is estimated by fitting a surface of the form  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$  through data on  $x_1, x_2$ , and  $y$ . Similarly for  $E[y|x_1, x_3]$ , etc.

## 4 Experimental Design

Experimental design refers to a collection of mathematical and empirical findings compiled to provide guidance on how to avoid some pitfalls in research, and how to obtain the most accurate and precise estimates possible, given finite data.<sup>3</sup> As shown in equation (1), variability of the response can be decomposed into two parts, “explained” and “unexplained.” The former measures the portion of the total variability which can be attributed to some predictor/factor, while the latter is simply the remaining variability, generally attributed to errors that cannot be accounted for. In a real experiment involving natural processes, due to the existence of observational errors, the unexplained component is not zero, and so, many experimental design methods are designed to minimize it. To that end, different sampling schemes of the experimental region are considered (among other things) to minimize the unexplained error. By contrast, in a computer experiment, all of the variability of the response is due to variability in the factors. In other words, there is no experimental error to minimize. For example, If the number of inputs is three, then  $V[y|x_1, x_2, x_3] = 0$ . In general, for computer data one has  $V[y|x_1, x_2, \dots, x_n] = 0$ , where  $n$  is the number

---

<sup>3</sup>Accuracy refers to how close an estimate is to the true value, while precision measures the variability of an estimate.

of inputs. The variability in  $y$  that is utilized to define all the sensitivity measures in Table 1 originates from the variability of the inputs that are *not* fixed. Either way, there *is* variability in  $y$  even for computer data, and so care must be taken in sampling, because, as shown below, the precision of the sensitivity measures depends on the choice of the sampling scheme.

One may wonder why sampling is necessary at all. After all, one can systematically vary each and every input from its minimum to its maximum, in some increment, thereby generating a lattice/grid that “samples” the experimental region. And the grid can be as fine as desired. As mentioned above, the problem with this scheme is that it can be computationally prohibitive; recall that each point in the sampling region requires running the  $\eta(\cdot)$  model, and this can be computationally prohibitive. For example, consider a  $n$ -dimensional lattice covering the  $n$ -dimensional space spanned by the  $n$  inputs. If each input takes 10 values (say), then  $10^n$  runs are required. Even for three inputs (such as in one of the examples considered here), that is 1000 runs of the model. Additionally, a grid does not lend itself to sequential/supplemental sampling. For example, if the 1000 points sampled from the experimental region are later found to be insufficient, there is no natural way of randomly choosing more points.

The alternative to a grid is random sampling from the experimental region. There exists a wide variety of random sampling techniques (Douglas 2005), but only two are considered here. The first technique is one of the simplest, called *simple random sampling*, which is random sampling without replacement. The second scheme is more sophisticated, and belongs to a class of sampling methods called *space-filling*;

the specific member used here is called *latin hypercube sampling* (Cioppa and Lucas 2007; Fang, Li, and Sudjianto 2006; Santner, Williams, and Notz 2003).

Another reason why a grid may be inadequate is that it is entirely possible that the inputs are in fact not independent. For example, it may be that in their natural environments, when input 1 is high, input 2 is low. In such a case, the situation where both inputs are high does not occur in the operational stage of the model; and yet, the grid includes those conditions in the experimental region. To avoid this problem, one treats the inputs as random variables following some specified probability distribution (e.g., Oakley and O'Hagan 2003). Then, random sampling should be done from that distribution.

## 4.1 Simple Random vs. Latin Hypercube Sampling

As mention previously, it is unfeasible to examine all possible values of all the inputs, and so, one resorts to some sampling scheme. One of the simplest is Simple Random Sampling (SRS), which has the desirable property of leading to the most precise estimate of the mean, if the population is homogeneous. For example, in sampling a continuous quantity such as height, if the population of heights has no clusters that distinguish different height characteristics, then a mean height computed from an SRS has the smallest variance across multiple samples (if multiple samples were taken). By contrast, when the population is not homogeneous throughout, consisting of (homogeneous) clusters/strata, then two alternatives are cluster sampling and

stratified sampling (Douglas 2005). Depending on the size of the clusters/strata and their number, one of these is guaranteed to produce more precise estimates of the mean than SRS.

In the above example, non-homogeneity of the population was implied to originate from some other *discrete* variable, the different levels of which correspond to the various clusters/strata in the population. In the situation, where the non-homogeneity is caused by a continuous quantity, there exist alternative sampling schemes which are again designed to yield more precise estimates. One class of such a scheme is called space-filling. The name is descriptive of the sampling scheme, and highlights the fact the simple random samples have a tendency to cluster. This clustering of data is not of concern when the population is homogeneous, but otherwise results in suboptimal estimates. Space-filling samples tend to scatter the sample across the full experimental region. One specific type of such a sampling scheme is called Latin Hypercube Sampling (LHS) (Cioppa 2007; Fang, Li, and Sudjianto 2006; Santner, Williams, and Notz 2003), which is best illustrated with an example.

Consider three discrete inputs, each with four levels. Then, the experimental region consists of  $4^3$  points. Now, denote the levels of first two inputs with the integers 1, 2, 3, 4, and let the letter, A, B, C, and D label the levels of the third input. Table 1 shows a sample of  $4^2$  points. For example, the top/left element of the table denotes the case where the first two inputs take the value 1, and the third input is at level A. The distinguishing characteristic of this table is that no two rows or columns contain more than one occurrence of the levels of the third input. This

table is an example of a latin square (Douglas 2005). A simple method for generating a latin square is to write down the elements of the first row, and then fill the next row by cyclic permutation of the previous row.

The latin square shown in Table 1 summarizes the design of an experiment. Note that in this design the  $\eta(\cdot)$  model is run only  $4^2$  times, a significant reduction compared to the size of the experimental region ( $4^3$ ). The next section shows that in spite of the relatively small size of such samples, not only the precision of many estimates is not hindered, but is in fact improved. The generalization to higher dimensions (i.e., more than three inputs) is, for obvious reasons, called Latin Hypercube Sampling (LHS). Variations on LHS have also been demonstrated to be beneficial in spatial statistics, where there exists a spatial correlation (Borkowski 2003).

## 5 A Simple Example

Figure 2 shows an example of a SRS (open circles) and a LHS (filled circles) taken from a 2-dimensional experimental region. This specific realization is uncharacteristic in that the SRS circles clearly cluster together, while the LHS circles do not. However, it serves to demonstrate the difference between the two sampling schemes. In short, SRS may lead to a “lumped” sample, but LHS cannot. Indeed, by design, no two cases in the LHS have the same values of  $x_1$  and  $x_2$ .

One may argue that the inability of SRS scheme to “fill” the experimental region is not of concern, because *in the long run* SRS samples will fill the space. It is true that

SRS samples fill the space, in the long run. But what that means is that an estimate based on an SRS sample is *accurate*, or more rigorously *unbiased*. If, however, one is interested in the *precision* of an estimate, then the two sampling schemes can lead to dramatically different results.

Consider the function

$$y = \eta(x_1, x_2) = x_1 + 50 x_2 + 2 x_1^2 + x_1 x_2 \quad , \quad (11)$$

shown in Figure 3. Note that the range of  $x_1$  is significantly wider than that of  $x_2$ ; (-2, 3) for the former, and (0.5, 1.0). This choice is intended to remind one that inputs often vary over different scales. First, suppose one is interested in the global mean of  $y$  (the  $z$ -axis). Of course, one can simply add all the  $y$  values of the data, and divide by the total number of cases, to arrive at the true mean. Now, consider taking a random sample from this data, computing the mean of the  $y$  values, and then repeating the procedure for another sample, some number of times. What is the distribution of the sample means for SRS and LHS? In other words, what is the empirical sampling distribution of the sample mean according to the two sampling schemes?

The top panel in Figure 4 shows the boxplots summarizing the empirical sampling distributions according to SRS (wide/black boxplots) and LHS (narrow/gray boxplots), for different sample sizes. The horizontal line marks the true mean. Evidently, both sampling schemes produce unbiased estimates of the mean, because all the boxplots cover the true mean. And not surprisingly, the variance of the distribu-

tions decreases with increasing sample size, as evident in the decrease in the height of the boxplots. But, somewhat surprisingly, the distributions for LHS are consistently smaller than the SRS distributions. This demonstrates that a LHS estimate of the mean is more precise than an estimate based on a SRS.

Intuitively, the reason this happens is that a given SRS may explore only the portion of the experimental region where the function takes low values. A different SRS may explore the high values of the function. Of course, on the average (i.e., in the long run), multiple SRS samples will examine all regions of the experimental region. But the existence of lumped samples leads to larger variability, as compared to the variability of the LHS scheme. Indeed, if the function had been  $y = \text{constant}$ , there would be no difference between the two sampling schemes.

In this example, although LHS leads to more precise estimates of the mean than SRS, it is important to point out that this is not true of all statistics. Figure 4 also shows the distributions of all the sensitivity measures shown in Table 2. It can be seen that for some measures (mean,  $V_{T1}$ ,  $V_{T2}$ ,  $S_2$ ,  $S_{T1}$ ) LHS does lead to more precise estimates, but for some measures ( $V_1$ ,  $V_2$ ,  $S_1$ ,  $S_{T2}$ ) the two schemes are comparable, (although one might still argue that  $V_{12}$  could belong to the former list.<sup>4</sup> Also note that many of these measures are biased for small sample sizes, because the boxplots do not sufficiently include the horizontal line.

---

<sup>4</sup>It can be shown that LHS cannot lead to less precise estimates than SRS. In fact, if the function  $\eta(\cdot)$  is monotonic, then this can be proved analytically (McKay, Beckman, and Conover 1979).

## 6 A More Complex Example - Lorenz '63

Although many of the results in the previous section can be computed analytically, the problem has been solved numerically in order to set the stage for the next, more realistic, example. This final example is one where the function representing the “black box” model is not available in analytic form - the Lorenz model (Lorenz 1963).

It is defined as

$$\begin{aligned}dX/dt &= -s(X - Y), \\dY/dt &= rX - Y - XZ, \\dZ/dt &= XY - bZ,\end{aligned}\tag{12}$$

where the model parameters are  $s$  (the Prandtl number),  $r$  (the Raleigh number), and  $b$ , the latter being a function of the wavenumber. The state space variables  $X$ ,  $Y$ , and  $Z$  measure the intensity of convective motion, and horizontal and vertical temperature variation, respectively (Bellomo and Preziosi 1995).

In terms of Figure 1, the Lorenz model parameters  $s$ ,  $r$ , and  $b$  are the inputs. In this paper, the outputs are not the state space variables  $X$ ,  $Y$ , and  $Z$ , but rather the maximum value of these quantities obtained over a 20-time-step forward integration of the Lorenz equations, in time-steps equal to 0.02. These numerical choices are mostly *ad hoc* and serve only to illustrate SA. Figure 5 shows each of these outputs as a function of the inputs for a wide range of input values. For each panel, the inputs that are not varied are set to their default (Lorenz 1963) values, 10, 28, and 8/3, for  $s$ ,  $r$ , and  $b$ , respectively; the input which is varied is incremented in steps leading to a

total of 50 equidistant values. The multiple curves correspond to 10 different initial conditions on the attractor.<sup>5</sup> It is important to note that in a realistic application of SA to a complex model, the inputs would not be varied in such a brute-force manner, because it would be computationally expensive. The only reason it is done here is to get an “inside look” at the “black box” in Figure 1.

It is well-known that the Lorenz model has different behavior for different model parameters, and this is reflected in the discontinuities apparent in Figure 5. It is unlikely that a realistic numerical model would have such discontinuities, and so, in order to simplify the illustration, the range of the model parameters is restricted only to regions where there are no discontinuities. The results are shown in Figure 6.

A brief description of the resulting computer data is in order. All of the relations are mostly linear. As such, the quadratic order of the polynomial regression emulating the expected values is sufficient. Moreover, it is evident that  $X_{max}$  generally increases with  $s$ ,  $r$ , or  $b$ , but at different rates (i.e., slope), and with different strengths (i.e., correlation).  $Y_{max}$  differs from  $X_{max}$  with respect to its dependence on  $s$ ; whereas  $X_{max}$  increases with  $s$ ,  $Y_{max}$  decreases with  $s$ .  $Z_{max}$  behaves similarly to  $Y_{max}$ , except that the strength of the relations (i.e. correlation) is stronger for  $Z_{max}$ .

As mentioned above, the data for performing SA is not obtained by systematically

---

<sup>5</sup>The initial conditions are obtained as follows: First, a random initial condition is selected, and the Lorenz model is integrated forward for a large number (500) of small time-steps (0.05). This locus of  $X, Y, Z$  points defines the attractor. Then random points are selected from this attractor, to serve as initial conditions for further runs.

incrementing all the inputs, for that would be computationally prohibitive for most realistic models. Thus, even though the Lorenz model is sufficiently simple to allow a brute-force approach, the SA is based on the two random sampling approaches - SRS, and LHS. There exists, however, one additional issue that arises in taking random samples. Technically, the random sample should be taken from the experimental region spanned by the parameters  $s$ ,  $r$ , and  $b$ . The resulting boxplots displaying the variability of the various sensitivity measures would then reflect sampling variability (see, e.g. Figure 4). However, there exists another source of variability in the Lorenz model, namely initial conditions. In other words, it is possible that different regions of the attractor have different sensitivity to the inputs. It would be possible to produce sensitivity results for the different regions of the attractor. But, here, for the sake of simplicity, and remaining focused on the illustration of variance-based SA methods, every time a sample is taken from the experimental region, the initial conditions are also varied. Consequently, the variability of the sensitivity measures is due to both sampling variability and initial conditions. This is the appropriate choice if one is interested in a “global” assessment of sensitivity across the entire attractor. (However, see Figure 8 and the corresponding discussion.)

Figure 7 shows the sensitivity measures for  $X_{max}$ ; results for  $Y_{max}$  and  $Z_{max}$  are not shown here. The boxplots summarize the corresponding distributions resulting from 200 trials (i.e., 200 different values of the model parameters and initial conditions). The top/left panel shows the  $V$  measures for  $s$ ,  $r$ , and  $b$ , from SRS (wide/black boxplots) and LHS (narrow/gray boxplots). One conclusion that follows from this is

that  $r$  is the more important of the three inputs in terms of how they affect  $X_{max}$ . The vicinity of the boxplot for  $s$  to the  $V = 0$  line suggests that sensitivity to  $s$  may not be statistically significant.

However, it would be a mistake to dismiss  $s$  as an important input, unless it has no interaction with the other parameters. The measures  $(V_{sr}, V_{sb}, V_{rb})$  must then be consulted; they are shown in the top/right panel. All three are nonzero. In particular,  $s$  appears to be an important parameter, by virtue of interaction with the other parameters. The lack of significant overlap between the boxplots for  $V_{sr}$  and  $V_{sb}$ , and the relative position of two, suggests that  $s$  interacts more with  $r$  than with  $b$ . By contrast, given the significant overlap of the boxplots for  $V_{sr}$  and  $V_{rb}$ , their equality cannot be ruled out. It is worth pointing out that even though the relative importance of the inputs as conveyed in top/left panel of Figure 7 may have been anticipated from Figure 6, the latter do not account for interactions. As such, the interaction measures shown in the top/right panel of Figure 7 convey information beyond a univariate analysis of the inputs.<sup>6</sup>

The remaining panels in Figure 7 convey similar results. Specifically, the pattern of the  $V$  measures is replicated in the  $V_T, S$ , and  $S_T$  measures. Note that the  $S_T$  measure for the  $s$  parameter is not zero; this is a consequence of the aforementioned conclusion that  $s$  cannot be ignored, even though it is not a terribly important input

---

<sup>6</sup>One may be concerned that the comparison of boxplots may not be valid, because of the paired nature of the design. As such, one should technically examine the boxplot of the *differences*, e.g.,  $V_{sr} - V_{sb}$ . We have confirmed that the conclusions are unaffected.

by itself.

Although most of the comparisons thus far have been based on the *relative* vertical position of the boxplots, the magnitude of the numbers is also useful. For example, according to the boxplot for  $S_T$  (i.e. middle boxplot, last panel), when  $s$  and  $b$  are fixed, about 70% of the variance in  $y$  can be attributed to  $r$ . If an interval estimate is required, one can report an inter-quartile range of 60% to 80%.

Finally, note that there is no appreciable difference between the two sampling schemes. Although, there is no unique explanation for that finding, one explanation is that variability is dominated by the variability from the initial conditions. Indeed, if one were to fix the initial condition to the default values  $(X, Y, Z) = (-14, -13, 47)$ , then the LHS does produce more precise estimates than SRS for several of the measures. These are shown in Figure 8, where the boxplots for SRS (wide/black) are generally longer than those for LHS (narrow/gray) for some of the measures. The difference between the two schemes is marginal, but it adds some support to the suspicion that variability due to initial conditions may play a role in assessing parameter importance in the Lorenz '63 model.

## 7 Conclusions and Discussion

The practical utility and intuitive appeal of Sensitivity Analysis (SA) has made it a popular tool in a wide range of fields. But in meteorological circles it usually refers to a set of anecdotal experiments, with little emphasis placed on statistical methods.

Recent work has elevated SA to a rigorous and active topic of research, drawing on methods of experimental design, analysis of variance, and regression and classification modeling (Fang, Li, and Sudjianto 2006; Santner, Williams, and Notz 2003). Here, one of the multitude of SA methods is described and demonstrated on a range of examples, from a toy model to the Lorenz ‘63 model. An application to a mature numerical weather prediction model (i.e., COAMPS<sup>®</sup>)<sup>7</sup> is currently in progress.

The specific method illustrated here is global, because it assesses the sensitivity of the output to inputs across the full range of inputs. By contrast, local methods rely on a derivative, inherently a local quantity, which addresses sensitivity only in a small neighborhood of a specific value of inputs. The method is also variance-based in that it examines how the variance (uncertainty) of the outputs is apportioned among the inputs. These two features make the methodology versatile and transparent at the same time. It can be applied in a wide range of problems, regardless of the complexity of the underlying relationships, and yet it provides simple, intuitive explanations of the results in terms of the decomposition of output variance. Although variance-based SA methods can also be used for ranking the inputs in order of their importance, they can go beyond ranking only, for example by providing guidance on which of the top-ranking inputs to keep. This is possible in variance based method, because one then knows what fraction of the total variance is explained by the surviving inputs.

In addition to the analysis of variance, two other ingredients of the method involve  
1) statistical emulation of data via regression methods, for the estimation of condi-

---

<sup>7</sup>COAMPS is a registered trademark of the Naval Research Laboratory.

tional expectations, and 2) experimental design (i.e. sampling) methods assuring that the sensitivity results are unbiased and optimally precise.

Several variance-based measures of sensitivity are considered. It is shown that some reduce to previously considered measures (e.g., standardized regression coefficients) when the inputs are independent. In general, however, regression coefficients are inadequate measures of sensitivity. It is also shown that measures have different properties with respect to two different sampling schemes - simple random sampling, and latin hypercube sampling. In the examples considered, the latter produces no less precise estimates than the former. This turns out to be a generally true comparison regarding the two sampling schemes. For this reason, it is recommended that latin hypercube sampling be considered instead of simple random sampling.

A few technical comments about the analysis performed here are in order. The boxplots in Figures 7 and 8 (and 4) have been employed to convey the variability of the respective measures. Comparison of the relative vertical position of the boxplots allows for a qualitative assessment of the importance of the corresponding inputs. However, such comparisons should be done with care. Quite generally, if boxplots for two variables (e.g., either for different inputs, or for different measures) have little or no overlap, then one can conclude that there is evidence for a difference between the two variables. However, if there is a significant overlap, one cannot conclude that the corresponding variables are not different. The reason is in the design of the experiment performed here; the data across the two boxplots are paired (Bailey 2008; Douglas 2005), because one set of 200 samples are employed for computing all of the

boxplots in the figures (for a given sampling scheme). The comparison across the two sampling schemes does not involve paired data, because the 200 samples for SRS are independent of the 200 samples used for LHS. In situations dealing with paired data, instead of comparing two boxplots (corresponding to two groups), one should examine a single boxplot of the differences between the two groups. In the examples considered here, the final conclusion of the analysis has been independent of how the comparisons are performed.

Also, here it has been sufficient to emulate the conditional expectations with a second-order polynomial regression. The main reason (for the sufficiency) is that the number of cases generated for “training” the regression model is practically unbounded, and so there is little or no chance of overfitting. The order of the polynomial has been fixed at two, because higher orders have not affected the results. In more realistic examples, where 1) the nonlinearity of the  $\eta(\cdot)$  model may require more nonlinear functions, and 2) the  $\eta(\cdot)$  model is expensive to run, alternative emulators should be used. One class of emulators which has gained recent popularity is called gaussian processes regression (Chaloner and Verdinelli 1995; Hsieh 2009; Sacks, Schiller, and Welch 1989; Sacks, et al. 1989; Welch et al. 1992). These are sophisticated models, whose training requires more effort than polynomial regression, which is the reason why they are not used here.

Finally, the analysis performed here has involved only one of the outputs ( $X_{max}$ ), and the approach can be applied to each output, separately; however, others have proposed multivariate methods where any covariance structure existing across the

outputs may also be taken into account (Fasso 2006; Oakley and O'Hagan 2004). This is a promising approach which will be examined in the context of ongoing work with COAMPS<sup>®</sup>,

### **Acknowledgments**

Scott Sandgathe is acknowledged for valuable discussion. Partial support for this project was provided by the Office of Naval Research grant numbers N00014-01-G-0460/0049 and N00014-05-1-0843.

## References

- Bailey, R.A., 2008: *Design of comparative experiments*, Cambridge series in Statistical and Probabilistic Mathematics (no. 25), Cambridge University Press, 346 pp.
- Bellomo, N., and L. Preziosi, 1995: *Modelling Mathematical Methods and Scientific Computation*. CRC press. 497 pp.
- Bishop, C. M., 1996: *Neural networks for pattern recognition*. Clarendon Press, Oxford, pp. 482.
- Borkowski, J. J., 2003: Simple Latin Square Sampling  $\pm k$  Designs. *Communications in Statistics, (Theory and Methods)*, **32(1)**, 215-237.
- Bowman, K. P., J. Sacks, and Y-F. Chang, 1993: Design and Analysis of Numerical Experiments. *J. Atmos. Sci.*, **50(9)**, 1267-1278.
- Bolado-Lavin, R., and A. C. Badea, 2008: Review of sensitivity analysis methods and experience for geological disposal of radioactive waste and spent nuclear fuel. JRC Scientific and Technical Report. Available online.
- Cacuci, D. G., 2003: *Sensitivity and Uncertainty Analysis*. Chapman and Hall/CRC. 285 pp.
- Chaloner, K., and I. Verdinelli, 1995: Bayesian experimental design: A review. *Statistical Science*, **10(3)**, 273-304.

- Cioppa, T. and T. Lucas, 2007: Efficient nearly orthogonal and space-filling latin hypercubes. *Technometrics*, **49(1)**, 4555.
- Douglas, C. M., 2005: *Design and Analysis of Experiments*, John Wiley & Sons, 643 pp.
- Draper, N. R., and H. Smith 1998, *Applied Regression Analysis*, Third Edition. John Wiley and Sons, Inc.
- Dusby, D., 2009: Hierarchical adaptive experimental design for Gaussian process emulators. *Reliability Engineering and System Safety*, **94(7)**,1183-1193 .
- Fang K.-T., R. Li, and A. Sudjianto, 2006: *Design and Modeling for Computer Experiments*, Chapman & Hall/CRC, 290 pp.
- Fasso, A. 2006: Sensitivity Analysis for Environmental Models and Monitoring Networks. In: Voinov, A., Jakeman, A.J., Rizzoli, A.E. (eds). Proceedings of the iEMSs Third Biennial Meeting: Summit on Environmental Modelling and Software. International Environmental Modelling and Software Society, Burlington, USA, July 2006.
- Internet: <http://www.iemss.org/iemss2006/sessions/all.html>
- ISBN 1-4243-0852-6 978-1-4243-0852-1
- Hsieh, W. 2009: *Machine Learning Methods in the Environmental Sciences: Neural Network and Kernels*, Cambridge University Press. 349 pp.
- Lorenz, E. N., 1963: Deterministic non-periodic flow. *J. Atmos. Sci.*, 20, 130-141.

- Jimenez, L., and D. Landgrebe, 1998: Supervised Classification in High Dimensional Space: Geometrical, Statistical, and Asymptotical Properties of Multivariate Data. *IEEE Transactions on System, Man, and Cybernetics*, **28**, Part C, 39-54. 10.1109/5326.661089
- McKay, M. D., R. J. Beckman, W. J., Conover, 1979: A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics*, **21(2)**, 239-245 .
- Miller, R. N., 2007: *Numerical Modeling of Ocean Circulation*. Cambridge University Press. 242 pp.
- Morris, M.D., 1991: Factorial sampling plans for preliminary computational experiments. *Technometrics*, **33**, 161174.
- Oakley, J. E., and A. O'Hagan, 2004: Probabilistic sensitivity analysis of complex models: a Bayesian approach. *J. R. Statist. Soc.*, **B 66(3)**, 751-769.
- Richardson, L. F. 2007: *Weather Prediction by Numerical Process (2nd Edition)*. Cambridge University Press. 250 pp.
- Rushton, K.R., 2003: *Groundwater Hydrology: Conceptual and Computational Models*. John Wiley and Sons Ltd. 416 pp. ISBN 0-470-85004-3
- Sacks, J., S. B. Schiller, and W. J. Welch, 1989: Designs for Computer Experiments. *Technometrics*, **31(1)**, 41-47.

- Sacks, J., W. J., Welch, T. J. Mitchell, H. P. Wynn, 1989: Design and Analysis of Computer Experiments. *Statistical Science*, **4**, 409-423.
- Saltelli, A., P. Annoni, I. Azzini, F. Campolongo, M. Ratto, S. Tarantola, 2010: Variance based sensitivity analysis of model output: Design and estimator for the total sensitivity index. *Computer Physics Communications*, **181**, 259270.
- Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, M. Saisana, S. Tarantola, 2008: *Global Sensitivity Analysis: The Primer*, Wiley Publishing, 304 pp.
- Santner, T.J., B. J. Williams, and W. I. Notz, 2003: *The Design and Analysis of Computer Experiments*. Springer, 299pp.
- Scott, D. W., 1992: *Multivariate Density Estimation; Theory, Practice, and Visualization*. New York: John Wiley & Sons, 325 pp.
- Sobol', I. M. 1993: Sensitivity estimates for nonlinear mathematical models, *Mathematical Modeling and Computational Experiments*, v. 1, 407-414.
- Weiss, N.A., 2005: *A Course in Probability*. AddisonWesley. 789 pp.
- Welch, W. J., R. J. Buck, J. Sacks, H. P. Wynn, T. J. Mitchell, and M. D. Morris, 1992: Screening, Predicting, and Computer Experiments. *Technometrics*, **34(1)**, 15-25.

## Figure Captions

Figure 1. A general representation of a complex model with inputs  $x_i, i = 1, \dots, n$ , (e.g., model parameters), and outputs  $y_j, j = 1, \dots, m$  (e.g., forecast parameters).

Figure 2. A SRS (open circles) and a LHS (filled circles). The former is clustered, while the latter “fills” the space.

Figure 3. An example of  $y = \eta(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$ .

Figure 4. Top left panel: Distribution of means according to SRS (wide/black boxplots) and LHS (narrow/gray boxplots), for different sample sizes, based on 1000 trials. The remaining panels show the distribution of the sensitivity measures in Table 1. The horizontal lines display the true value of the respective quantity in each panel.

Figure 5. The dependence of the outputs of the Lorenz '63 models ( $X_{max}, Y_{max}, Z_{max}$ ) on its inputs ( $s, r, b$ ). In each panel only one of the inputs is varied, while the other two inputs are held at their respective default value. The 10 “curves” correspond to 10 different initial conditions.

Figure 6. Same as Figure 5, except zoomed in onto regions without discontinuities.

Figure 7. Variance-based sensitivity measures for the Lorenz '63 model. The boxplots show variability due to sampling variability in the experimental region and due to initial conditions.

Figure 8. Same as Figure 7, except the variability displayed in the boxplots is due

solely to sampling in the experimental region; the initial conditions are set to their default values.

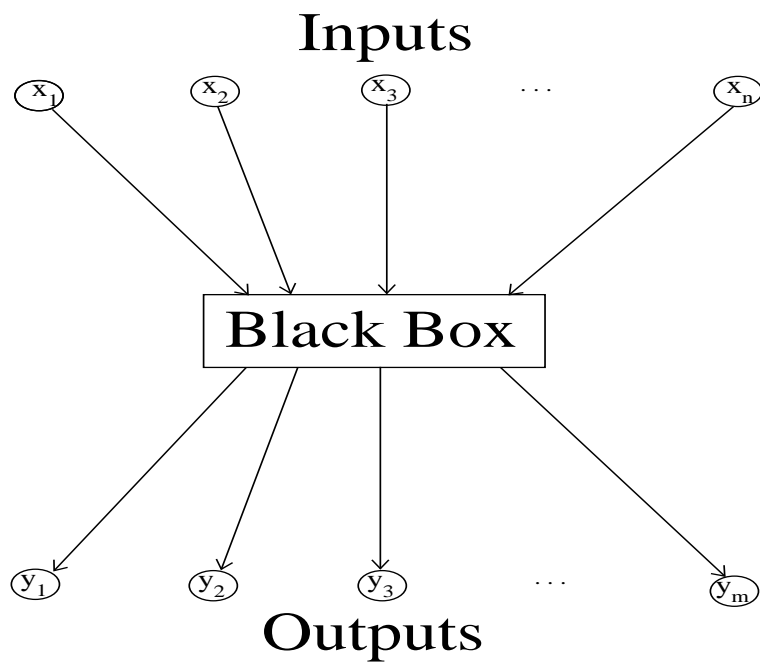


Figure 1. A general representation of a complex model with inputs  $x_i, i = 1, \dots, n$ , (e.g., model parameters), and outputs  $y_j, j = 1, \dots, m$  (e.g., forecast parameters).

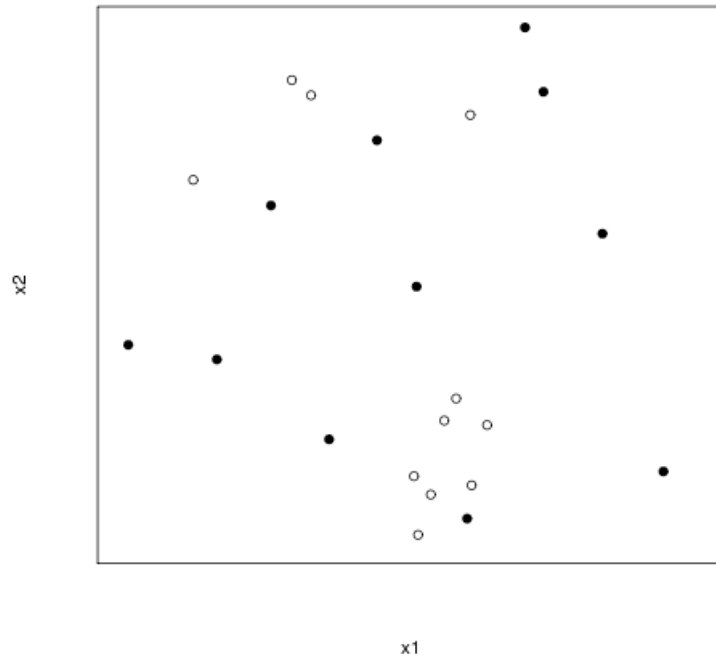


Figure 2. A SRS (open circles) and a LHS (filled circles). The former is clustered, while the latter “fills” the space.

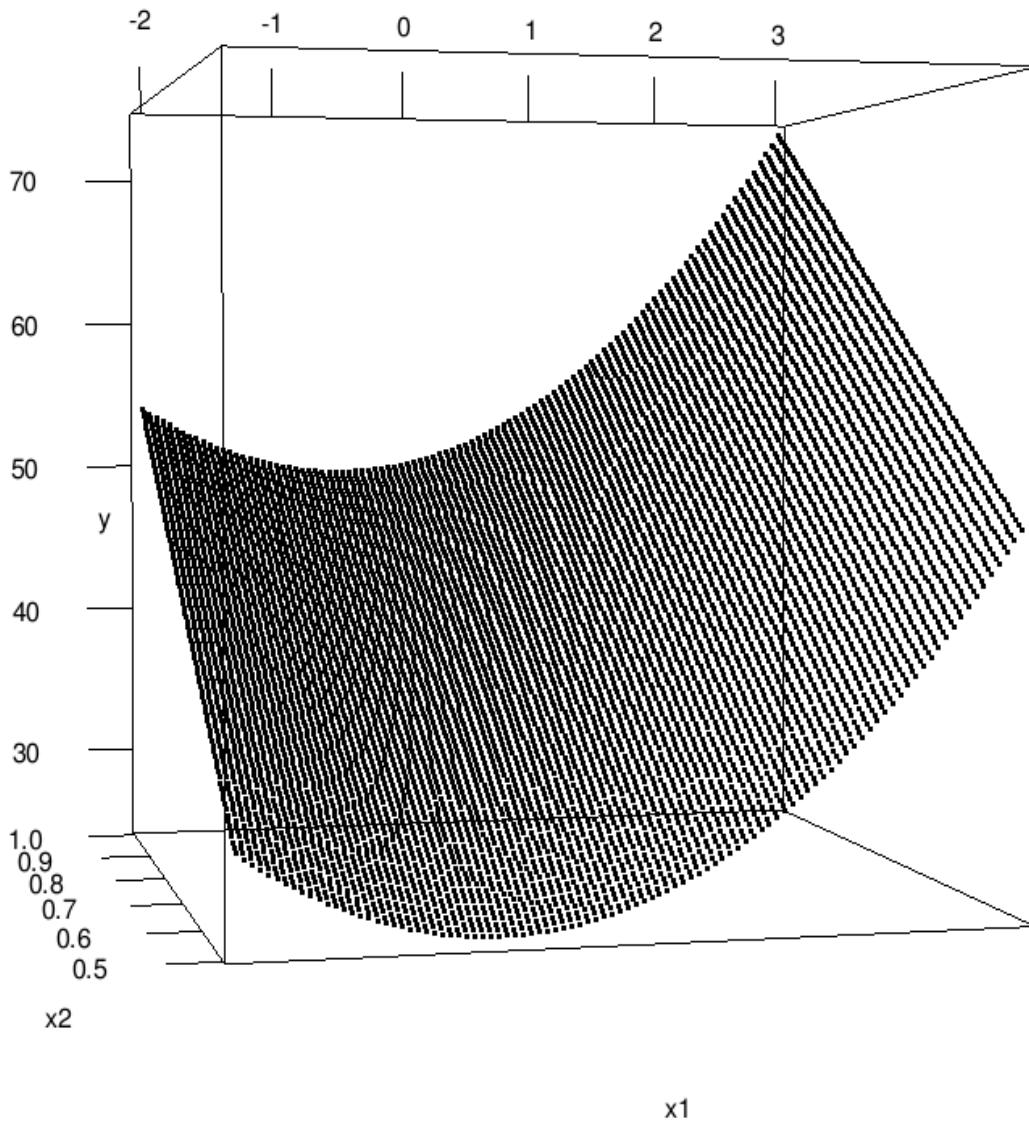


Figure 3. An example of  $y = \eta(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$ .

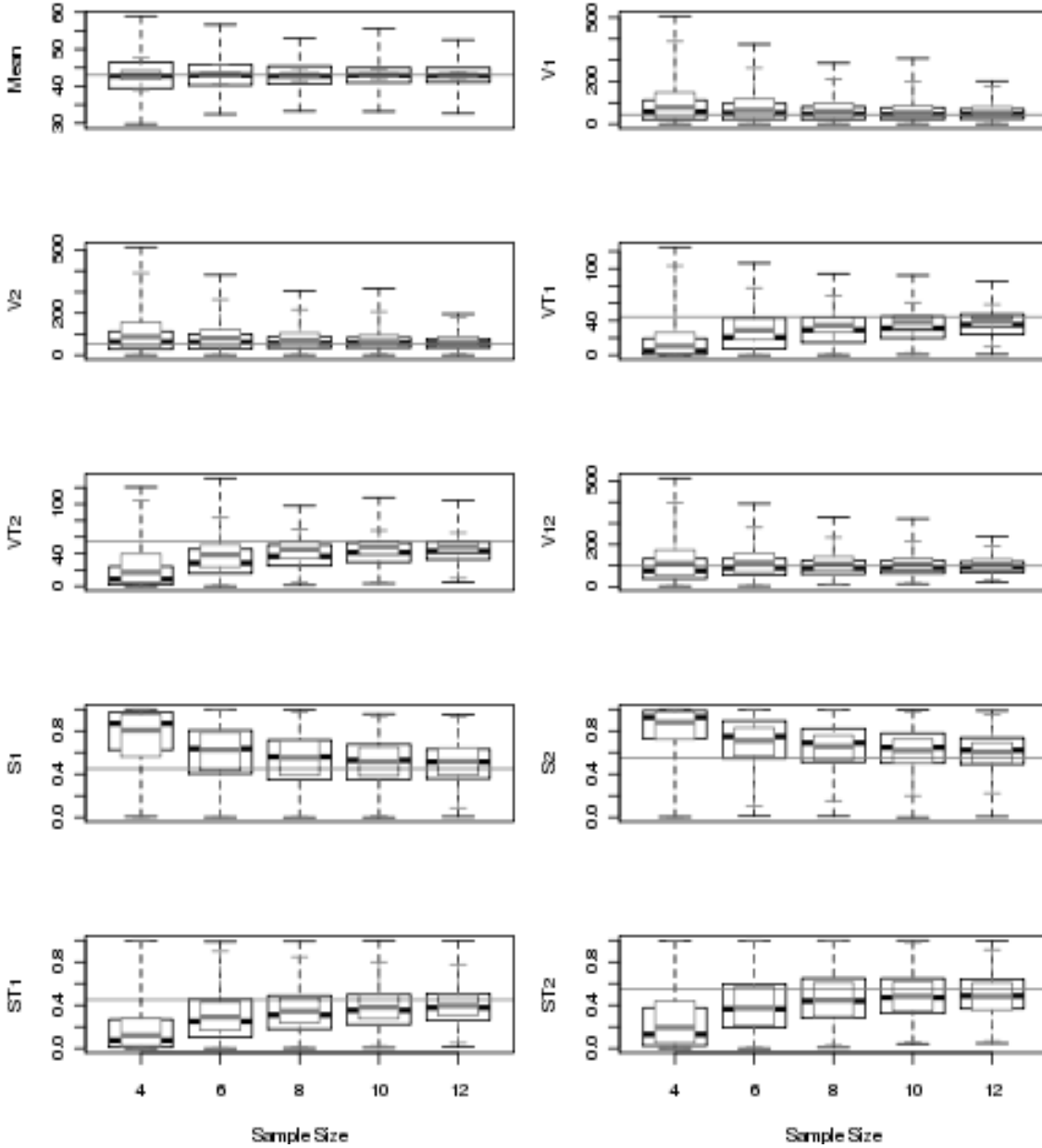


Figure 4. Top left panel: Distribution of means according to SRS (wide/black boxplots) and LHS (narrow/gray boxplots), for different sample sizes, based on 1000 trials. The remaining panels show the distribution of the sensitivity measures in Table 1. The horizontal lines display the true value of the respective quantity in each panel.

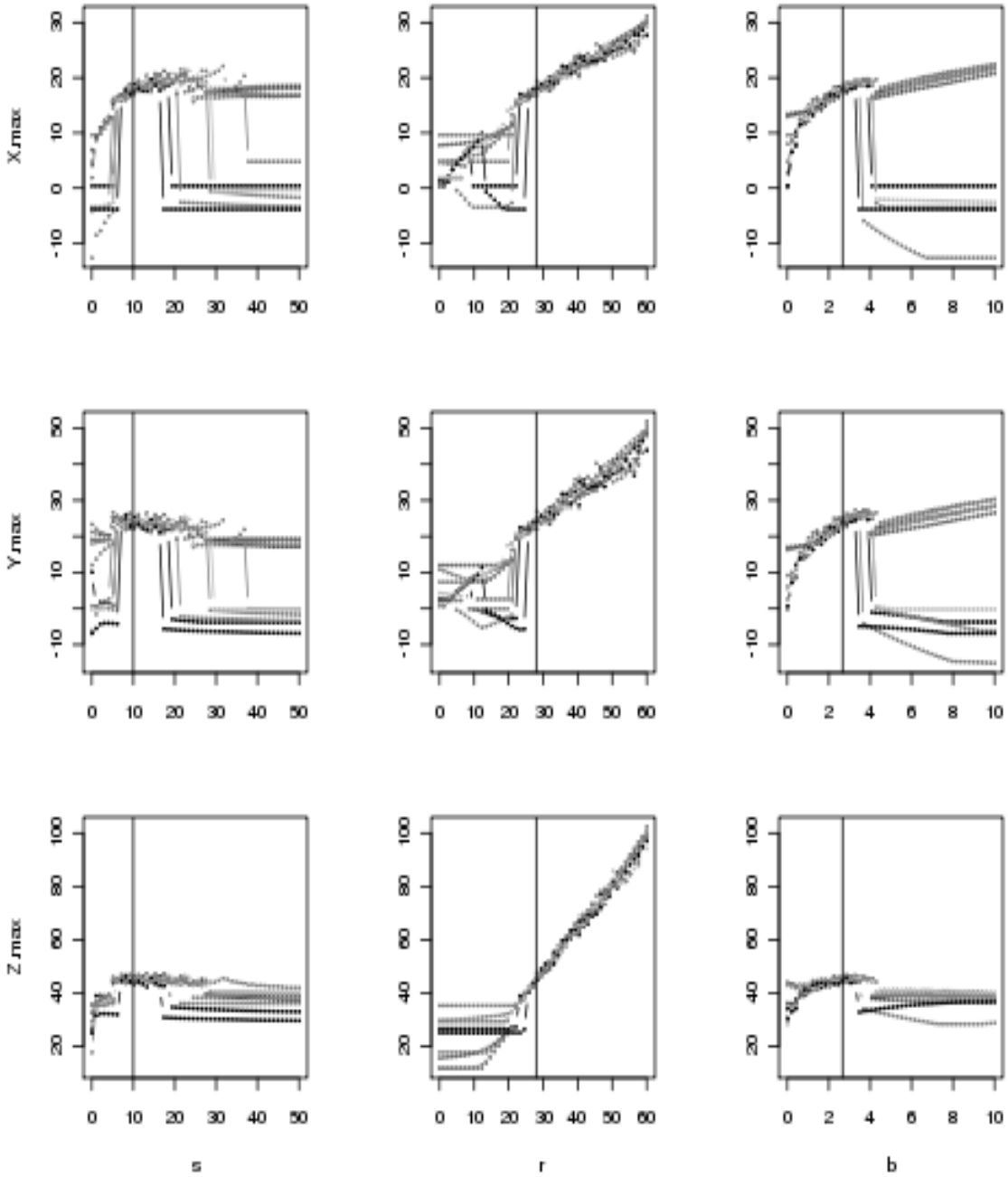


Figure 5. The dependence of the outputs of the Lorenz '63 models ( $X_{max}, Y_{max}, Z_{max}$ ) on its inputs ( $s, r, b$ ). In each panel only one of the inputs is varied, while the other two inputs are held at their respective default value. The 10 “curves” correspond to 10 different initial conditions.

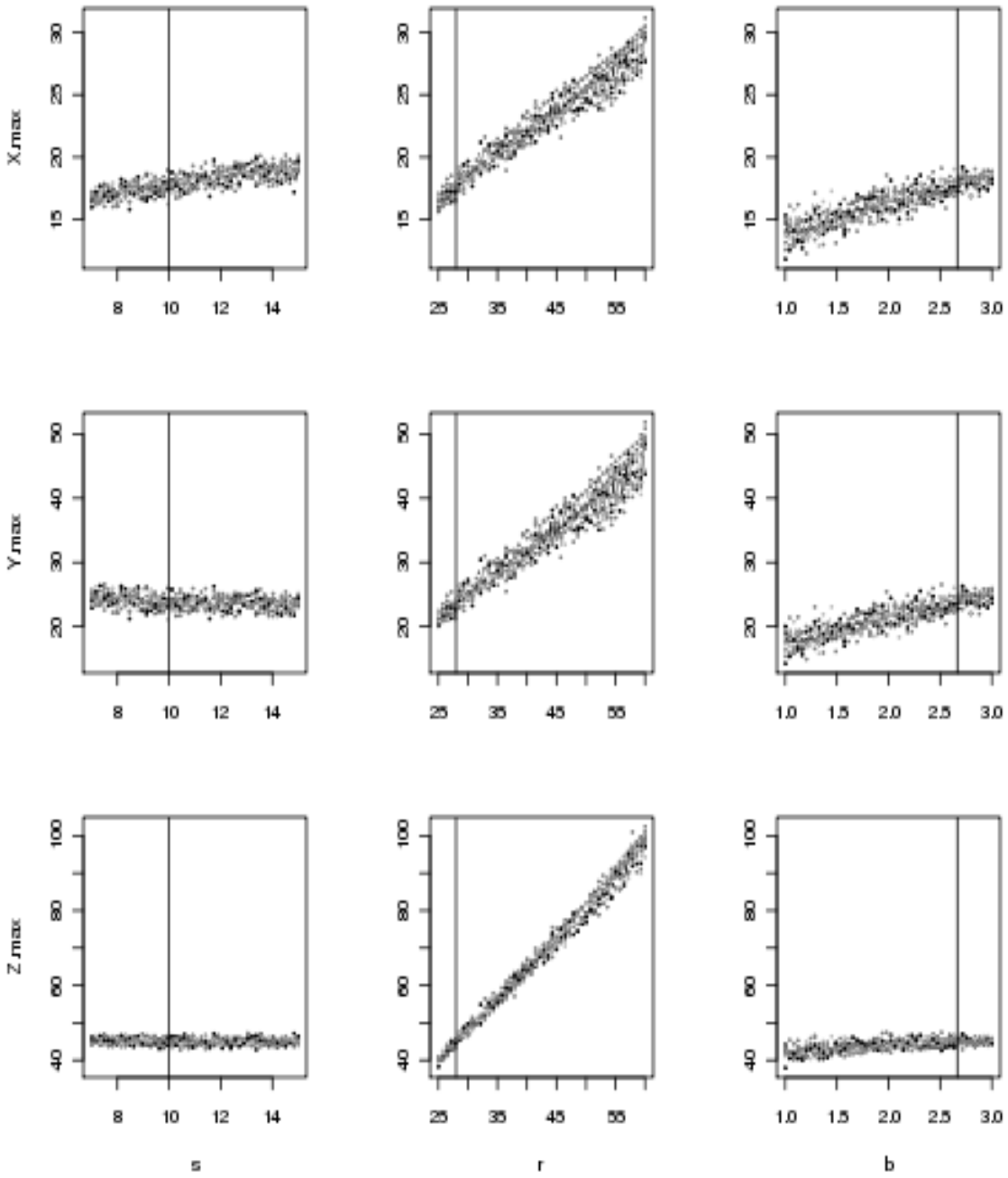


Figure 6. Same as Figure 5, except zoomed in onto regions without discontinuities.

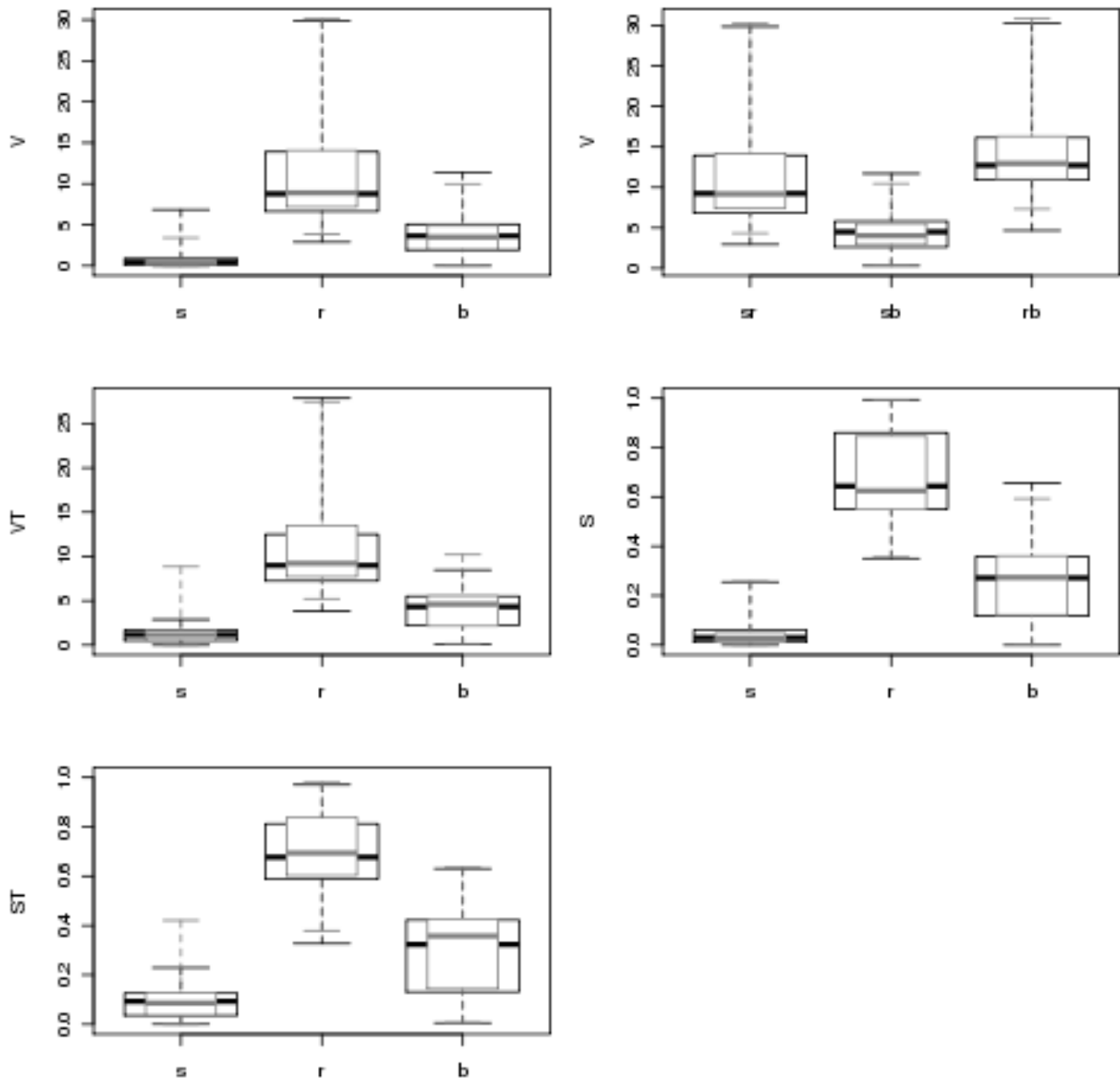


Figure 7. Variance-based sensitivity measures for the Lorenz '63 model. The boxplots show variability due to sampling variability in the experimental region and due to initial conditions.

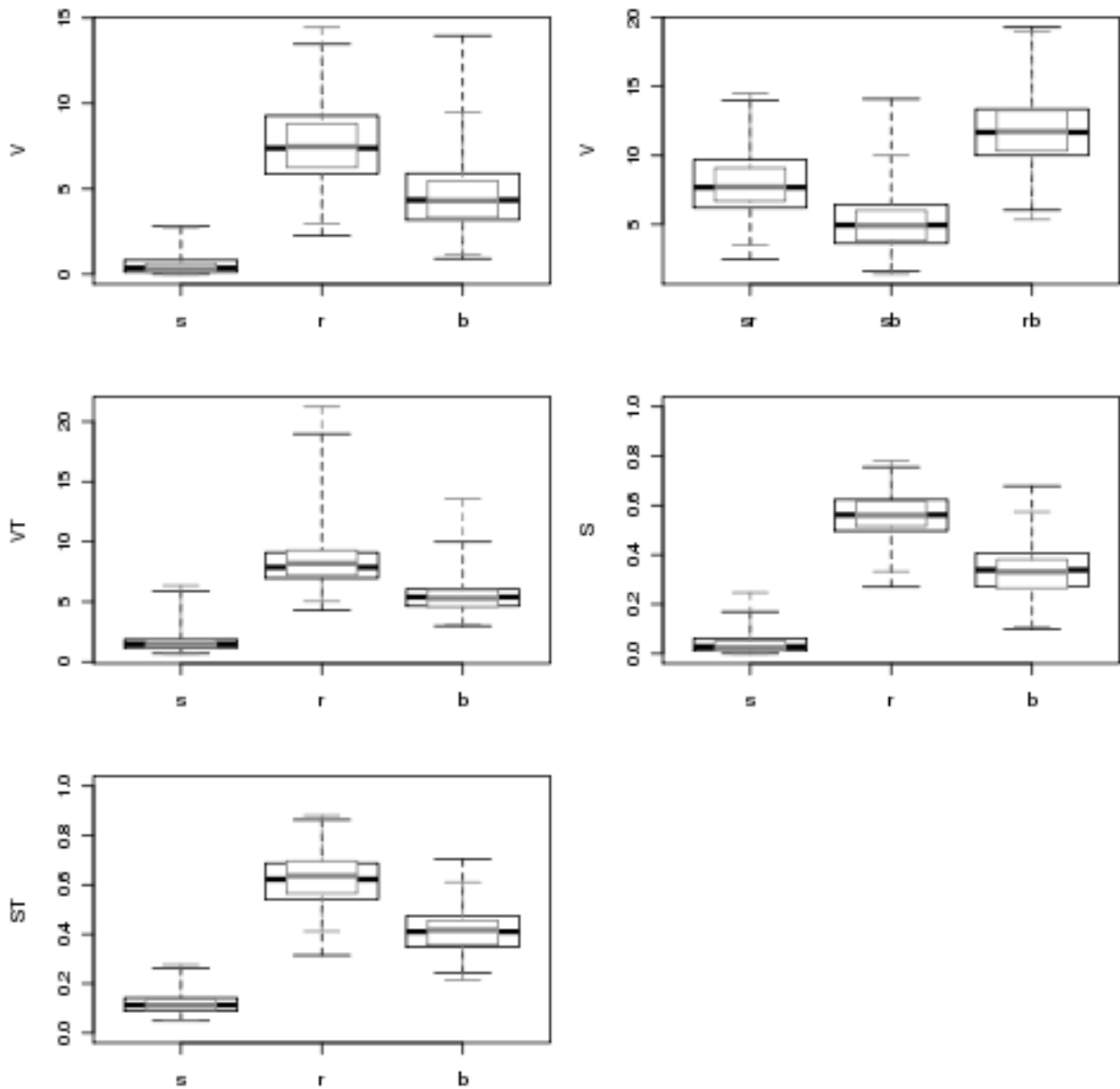


Figure 8. Same as Figure 7, except the variability displayed in the boxplots is due solely to sampling in the experimental region; the initial conditions are set to their default values.

Sensitivity measure	Meaning
$V_i = V[E[y x_i]]$	Reduction in uncertainty of $y$ , given $x_i$
$V_{ij} = V[E[y x_i, x_j]]$	Reduction in uncertainty of $y$ , given $x_i$ and $x_j$
$V_{T1} = V[y] - V[E[y x_2, x_3, \dots]]$	Uncertainty in $y$ remaining, given everything except $x_1$
$S_i = V_i/V[y]$	Main effect index of $x_i$
$S_{Ti} = V_{Ti}/V[y]$	Total effect index of $x_i$

Table 1. The sensitivity measures examined here.

measure	General	Indep $x_1, x_2$
$z_1$	$x_1 - E[x_1]$	$x_1 - E[x_1]$
$z_2$	$E[x_1 x_2] - E[x_1]$	0
$z_{12}$	$-z_2(x_2)$	0
$V_1$	$V[x_1]$	$V[x_1]$
$V_2$	$V[E[x_1 x_2]]$	0
$V_{12}$	$V[x_1]$	$V[x_1]$
$V_{T1}$	$V[x_1] - V_2$	$V[x_1]$
$V_{T2}$	0	0
$S_1$	1	1
$S_2$	$V_2/V[x_1]$	0
$S_{T1}$	$1 - S_2$	1
$S_{T2}$	0	0

Table 2. The sensitivity measures for the “black box” model  $y = \eta(x_1, x_2) = x_1$ .

measure	Indep $x_1, x_2$	$E[x_1] = E[x_2] = 0$
$z_1$	$(\beta_1 + \beta_{12}E[x_2]) (x_1 - E[x_1])$	$\beta_1 x_1$
$z_2$	$(\beta_2 + \beta_{12}E[x_1]) (x_2 - E[x_2])$	$\beta_2 x_2$
$z_{12}$	$\beta_{12} (x_1 - E[x_1]) (x_2 - E[x_2])$	$\beta_{12} x_1 x_2$
$V_1$	$(\beta_1 + \beta_{12}E[x_2])^2 V[x_1]$	$\beta_1^2 V[x_1]$
$V_2$	$(\beta_2 + \beta_{12}E[x_1])^2 V[x_2]$	$\beta_2^2 V[x_2]$
$V_{12}$	$V[y]$	$\beta_1^2 V[x_1] + \beta_2^2 V[x_2] + \beta_{12}^2 V[x_1]V[x_2]$
$V_{T1}$	$V[y] - (\beta_2 + \beta_{12}E[x_1])^2 V[x_2]$	$(\beta_1^2 + \beta_{12}^2 V[x_2])V[x_1]$
$V_{T2}$	$V[y] - (\beta_1 + \beta_{12}E[x_2])^2 V[x_1]$	$(\beta_2^2 + \beta_{12}^2 V[x_1])V[x_2]$
$S_1$	$(\beta_1 + \beta_{12}E[x_2])^2 V[x_1]/V[y]$	$\beta_1^2 V[x_1]/V[y]$
$S_2$	$(\beta_2 + \beta_{12}E[x_1])^2 V[x_2]/V[y]$	$\beta_2^2 V[x_2]/V[y]$
$S_{T1}$	$1 - S_2$	$1 - S_2$
$S_{T2}$	$1 - S_1$	$1 - S_1$

Table 3. The sensitivity measures for the “black box” model  $y = \eta(x_1, x_2) = x_1 +$

$50 x_2 + 2 x_1^2 + x_1 x_2$ .

	1	2	3	4
1	A	B	D	C
2	B	D	C	A
3	D	C	A	B
4	C	A	B	D

Table 4. An example of a latin square, with row and column representing two inputs with levels 1, 2, 3, and 4. The letters A, B, C, D denote the 4 levels of a third input.